Express Mail No.EL675506812US

Attorney Docket No. 3414

PATENT APPLICATION

COMPLEXITY MANAGEMENT OF GENOMIC DNA

INVENTORS

Xing Su, a citizen of the United States
residing at: 21811 Granada Avenue
Cupertino, CA 95014

Hajime Matsuzaki, a citizen of Japan
Residing at: 4078 Middlefield Road,
Palo Alto, CA 94303

Giulia Kennedy, a citizen of the United States
Residing at: 360 Castenada Avenue
San Francisco, CA 94116

ASSIGNEE

Affymetrix, Inc.
3380 Central Expressway
Santa Clara, CA 95051

PATENT

Attorney Docket No. 3414

FIELD OF THE INVENTION

The invention relates to enrichment and amplification of sequences from a nucleic acid sample. In one embodiment, the invention relates to enrichment and amplification of nucleic acids for the purpose of further analysis. The present invention relates to the fields of molecular biology and genetics.

BACKGROUND OF THE INVENTION

The past years have seen a dynamic change in the ability of science to comprehend vast amounts of data. Pioneering technologies such as nucleic acid arrays allow scientists to delve into the world of genetics in far greater detail than ever before. Exploration of genomic DNA has long been a dream of the scientific community. Held within the complex structures of genomic DNA lies the potential to identify, diagnose, or treat diseases like cancer, Alzheimer disease or alcoholism. Exploitation of genomic information from plants and animals may also provide answers to the world's food distribution problems.

Recent efforts in the scientific community, such as the publication of the draft sequence of the human genome in February 2001, have changed the dream of genome exploration into a reality. Genome-wide assays, however, must contend with the complexity of genomes; the human genome for example is estimated to have a complexity of $3x10^9$ base pairs. Novel methods of sample preparation and sample analysis that reduce complexity may provide for the fast and cost effective exploration of complex samples of nucleic acids, particularly genomic DNA.

SUMMARY OF THE INVENTION

The present invention provides for novel methods of sample preparation and analysis comprising managing or reducing the complexity of a nucleic acid sample. The methods of the invention generally involve controlling the average length of product in an amplification reaction by varying the conditions and or components of the reaction so that size selection and target amplification are achieved in a single step. The methods are

preferably non-gel based. For many of the embodiments the step of complexity reduction may be performed entirely in a single tube. The invention further provides for analysis of the sample by hybridization to an array, which may be specifically designed to interrogate fragments for particular characteristics, such as, for example, the presence or absence of a polymorphism. The invention further provides for methods of designing an array to interrogate particular subsets of fragments. In a preferred embodiment the invention discloses novel methods of genome-wide polymorphism discovery and genotyping.

In one embodiment the step of complexity management of the nucleic acid comprises fragmenting the nucleic acid sample to form fragments, ligating adaptor sequences to the fragments and amplifying the fragments under conditions that favor amplification of a particular size range of fragments.

Conditions that may be varied include: the extension time, the annealing time, concentration of primer, primer length, presence or absence of a 3' to 5' exonuclease activity and concentration of nucleotide analogs. Another step that can be used to control average length of the amplification product is the introduction of regions of complementarity in the 5' and 3' ends of the target fragments. One way to accomplish this is through ligation of a single adaptor sequence to both ends of the fragments. In general the methods use a single round of amplification, but in some embodiments the first amplification product is diluted and amplified with a second round of amplification which is preferably done under conditions that favor amplification of a particular size range of fragments.

In one embodiment, the invention relates to a kit comprising reagents and instructions for amplifying a subset of fragments. The kit may comprise reagents and instructions necessary for amplification of one or more subsets of fragments.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 is a chart of some parameters that can be altered to control the average size of the product in a PCR.

Figure 2 is a schematic representation of the effect of amplification of fragments with complementary ends.

Figure 3 shows the results of size fractionation by varied primer concentration.

Figure 4 shows how *in silico* digestion can be used to predict the size of restriction fragments containing SNPs.

Figure 5 is a table of the number of SNPs predicted to be found on 400 to 800 base pair fragments when genomic DNA is digested with the restriction enzyme in column 1.

Figure 6 is a flow chart showing design of an array in conjunction with size selection of SNP containing fragments.

Figure 7 shows the results of size selection PCR and hybridization of the product to an array interrogating for presence or absence of a subset of SNPs.

Figure 8 shows a schematic of size selection PCR using short annealing and extension times.

Figure 9 shows the results of hybridization of PCR products, generated using short annealing and extension times to amplify fragments of a selected size, to an array.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

(A) General

The present invention relies on many patents, applications and other references for details known to those of the art. Therefore, when a patent, application, or other reference is cited or repeated below, it should be understood that it is incorporated by reference in its entirety for all purposes as well as for the proposition that is recited. As used in the specification and claims, the singular form "a," "an," and "the" include plural references unless the context clearly dictates otherwise. For example, the term "an agent" includes a plurality of agents, including mixtures thereof. An individual is not limited to a human being but may also be other organisms including but not limited to mammals, plants, bacteria, or cells derived from any of the above.

Throughout this disclosure, various aspects of this invention are presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as common individual numerical values within that range. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within that range, for example, 1, 2, 3, 4, 5, and 6. The same holds true for ranges in increments of 10^5, 10^4, 10^3, 10^2, 10, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, or 10^{-5}, for example. This applies regardless of the breadth of the range.

The practice of the present invention may employ, unless otherwise indicated, conventional techniques of organic chemistry, polymer technology, molecular biology (including recombinant techniques), cell biology, biochemistry, and immunology, which are within the skill of the art. Such conventional techniques include polymer array synthesis, hybridization, ligation, and detection of hybridization using a label. Specific illustrations of suitable techniques can be had by reference to the example hereinbelow. However, other equivalent conventional procedures can, of course, also be used. Such conventional techniques can be found in standard laboratory manuals such as *Genome Analysis: A Laboratory Manual Series (Vols. I-IV)*, *Using Antibodies: A Laboratory Manual*, *Cells: A Laboratory Manual*, *PCR Primer: A Laboratory Manual*, and *Molecular Cloning: A Laboratory Manual* (all from Cold Spring Harbor Laboratory Press), all of which are herein incorporated in their entirety by reference for all purposes.

Some aspects of the present invention make use of microarrays, also called arrays. Methods and techniques applicable to array synthesis have been described in U.S. Patents Nos. 5,143,854, 5,242,974, 5,252,743, 5,324,633, 5,384,261, 5,424,186, 5,451,683, 5,482,867, 5,491,074, 5,527,681, 5,550,215, 5,571,639, 5,578,832, 5,593,839, 5,599,695, 5,624,711, 5,631,734, 5,795,716, 5,831,070, 5,837,832, 5,856,101, 5,858,659, 5,936,324, 5,968,740, 5,974,164, 5,981,185, 5,981,956, 6,025,601, 6,033,860, 6,040,193, and 6,090,555. All of the above patents incorporated herein by reference in their entireties for all purposes.

The word "DNA" may be used below as an example of a nucleic acid. It is understood that this term includes all nucleic acids, such as DNA and RNA, unless a use below requires a specific type of nucleic acid.

(B) Definitions

Nucleic acids according to the present invention may include any polymer or oligomer of pyrimidine and purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively. (*See* Albert L. Lehninger, *Principles of Biochemistry*, at 793-800 (Worth Pub. 1982) which is herein incorporated in its entirety for all purposes). Indeed, the present invention contemplates any deoxyribonucleotide, ribonucleotide or peptide nucleic acid component, and any chemical variants thereof, such as methylated, hydroxymethylated or glucosylated forms of these bases, and the like. The polymers or oligomers may be heterogeneous or homogeneous in composition, and may be isolated from naturally occurring sources or may be artificially or synthetically produced. In addition, the nucleic acids may be DNA or RNA, or a mixture thereof, and may exist permanently or transitionally in single-stranded or double-stranded form, including homoduplex, heteroduplex, and hybrid states.

An "oligonucleotide" or "polynucleotide" is a nucleic acid ranging from at least 2, preferably at least 8, 15 or 20 nucleotides in length, but may be up to 50, 100, 1000, or 5000 nucleotides long or a compound that specifically hybridizes to a polynucleotide. Polynucleotides of the present invention include sequences of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) or mimetics thereof which may be isolated from natural sources, recombinantly produced or artificially synthesized. A further example of a polynucleotide of the present invention may be a peptide nucleic acid (PNA). (*See* U.S. Patent No. 6,156,501 which is hereby incorporated by reference in its entirety.) The invention also encompasses situations in which there is a nontraditional base pairing such as Hoogsteen base pairing which has been identified in certain tRNA molecules and postulated to exist in a triple helix. "Polynucleotide" and "oligonucleotide" are used interchangeably in this application.

The term "fragment," "segment," or "DNA segment" refers to a portion of a larger DNA polynucleotide or DNA. A polynucleotide, for example, can be broken up,

or fragmented into, a plurality of segments. Various methods of fragmenting nucleic acid are well known in the art. These methods may be, for example, either chemical or physical in nature. Chemical fragmentation may include partial degradation with a DNase; partial depurination with acid; the use of restriction enzymes; intron-encoded endonucleases; DNA-based cleavage methods, such as triplex and hybrid formation methods, that rely on the specific hybridization of a nucleic acid segment to localize a cleavage agent to a specific location in the nucleic acid molecule; or other enzymes or compounds which cleave DNA at known or unknown locations. Physical fragmentation methods may involve subjecting the DNA to a high shear rate. High shear rates may be produced, for example, by moving DNA through a chamber or channel with pits or spikes, or forcing the DNA sample through a restricted size flow passage, e.g., an aperture having a cross sectional dimension in the micron or submicron scale. Other physical methods include sonication and nebulization. Combinations of physical and chemical fragmentation methods may likewise be employed such as fragmentation by heat and ion-mediated hydrolysis. *See* for example, Sambrook et al., "Molecular Cloning: A Laboratory Manual," 3rd Ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (2001) ("Sambrook et al.) which is incorporated herein by reference for all purposes. These methods can be optimized to digest a nucleic acid into fragments of a selected size range. Useful size ranges may be from 100, 200, 400, 700 or 1000 to 500, 800, 1500, 2000, 4000 or 10,000 base pairs. However, larger size ranges such as 4000, 10,000 or 20,000 to 10,000, 20,000 or 500,000 base pairs may also be useful.

A number of methods disclosed herein require the use of restriction enzymes to fragment the nucleic acid sample. In general, a restriction enzyme recognizes a specific nucleotide sequence of four to eight nucleotides and cuts the DNA at a site within or a specific distance from the recognition sequence. For example, the restriction enzyme *Eco*RI recognizes the sequence GAATTC and will cut a DNA molecule between the G and the first A. The length of the recognition sequence is roughly proportional to the frequency of occurrence of the site in the genome. A simplistic theoretical estimate is that a six base pair recognition sequence will occur once in every 4096 (4^6) base pairs while a four base pair recognition sequence will occur once every 256 (4^4) base pairs. *In silico* digestions of sequences from the Human Genome Project show that the actual

occurrences are even more infrequent, depending on the sequence of the restriction site. Because the restriction sites are rare, the appearance of shorter restriction fragments, for example those less than 1000 base pairs, is much less frequent than the appearance of longer fragments. Many different restriction enzymes are known and appropriate restriction enzymes can be selected for a desired result. (For a description of many restriction enzymes *see*, New England BioLabs Catalog which is herein incorporated by reference in its entirety for all purposes).

"Adaptor sequences" or "adaptors" are generally oligonucleotides of at least 5, 10, or 15 bases and preferably no more than 50 or 60 bases in length, however, they may be even longer, up to 100 or 200 bases. Adaptor sequences may be synthesized using any methods known to those of skill in the art. For the purposes of this invention they may, as options, comprise templates for PCR primers, restriction sites and promoters. The adaptor may be entirely or substantially double stranded. The adaptor may be phosphorylated or unphosphorylated on one or both strands. Adaptors are particularly useful in one embodiment of the current invention if they comprise a substantially double stranded region and short single stranded regions which are complementary to the single stranded region created by digestion with a restriction enzyme. For example, when DNA is digested with the restriction enzyme *Eco*RI the resulting double stranded fragments are flanked at either end by the single stranded overhang 5'-AATT-3', an adaptor that carries a single stranded overhang 5'-AATT-3' will hybridize to the fragment through complementarity between the overhanging regions. This "sticky end" hybridization of the adaptor to the fragment may facilitate ligation of the adaptor to the fragment but blunt ended ligation is also possible.

Adaptors can be used to introduce complementarity between the ends of a nucleic acid. For example, if a double stranded region of DNA is digested with a single enzyme so that each of the ends of the resulting fragments is generated by digestion with the same restriction enzyme, both ends will have the same overhanging sequence. For example if a nucleic acid sample is digested with *Eco*RI both strands of the DNA will have at their 5' ends a single stranded region, or overhang, of 5'-AATT-3'. A single adaptor that has a complementary overhang of 5'-AATT-3' can be ligated to both ends of the fragment. Each of the strands of the fragment will have one strand of the adaptor ligated to the 5'

end and the second strand of the adaptor ligated to the 3' end. The two strands of the adaptor are complementary to one another so the resulting ends of the individual strands of the fragment will be complementary.

A single adaptor can also be ligated to both ends of a fragment resulting from digestion with two different enzymes. For example, if the method of digestion generates blunt ended fragments, the same adaptor sequence can be ligated to both ends. Alternatively some pairs of enzymes leave identical overhanging sequences. For example, *Bgl*II recognizes the sequence 5'-AGATCT-3', cutting after the first A, and *Bam*HI recognizes the sequence 5'-GGATCC-3', cutting after the first G; both leave an overhang of 5'-GATC-3'. A single adaptor with an overhang of 5'-GATC-3' may be ligated to both digestion products.

Digestion with two or more enzymes can be used to selectively ligate separate adapters to either end of a restriction fragment. For example, if a fragment is the result of digestion with EcoRI at one end and BamHI at the other end, the overhangs will be 5'-AATT-3' and 5'GATC-3', respectively. An adaptor with an overhang of AATT will be preferentially ligated to one end while an adaptor with an overhang of GATC will be preferentially ligated to the second end.

Methods of ligation will be known to those of skill in the art and are described, for example in Sambrook et at. and the New England BioLabs catalog both of which are incorporated herein by reference for all purposes. Methods include using T4 DNA Ligase which catalyzes the formation of a phosphodiester bond between juxtaposed 5′ phosphate and 3′ hydroxyl termini in duplex DNA or RNA with blunt or and sticky ends; *Taq* DNA ligase which catalyzes the formation of a phosphodiester bond between juxtaposed 5′ phosphate and 3′ hydroxyl termini of two adjacent oligonucleotides which are hybridized to a complementary target DNA; *E.coli* DNA ligase which catalyzes the formation of a phosphodiester bond between juxtaposed 5′ -phosphate and 3′ -hydroxyl termini in duplex DNA containing cohesive ends; and T4 RNA ligase which catalyzes ligation of a 5′ phosphoryl-terminated nucleic acid donor to a 3′ hydroxyl-terminated nucleic acid acceptor through the formation of a 3′ –>5′ phosphodiester bond, substrates include single-stranded RNA and DNA as well as dinucleoside pyrophosphates; or any other methods described in the art.

"Genome" designates or denotes the complete, single-copy set of genetic instructions for an organism as coded into the DNA of the organism. A genome may be multi-chromosomal such that the DNA is cellularly distributed among a plurality of individual chromosomes. For example, in human there are 22 pairs of chromosomes plus a gender associated XX or XY pair.

The term "chromosome" refers to the heredity-bearing gene carrier of a living cell which is derived from chromatin and which comprises DNA and protein components (especially histones). The conventional internationally recognized individual human genome chromosome numbering system is employed herein. The size of an individual chromosome can vary from one type to another with a given multi-chromosomal genome and from one genome to another. In the case of the human genome, the entire DNA mass of a given chromosome is usually greater than about 100,000,000 bp. For example, the size of the entire human genome is about 3×10^9 bp. The largest chromosome, chromosome no. 1, contains about 2.4×10^8 bp while the smallest chromosome, chromosome no. 22, contains about 5.3×10^7 bp.

A "chromosomal region" is a portion of a chromosome. The actual physical size or extent of any individual chromosomal region can vary greatly. The term "region" is not necessarily definitive of a particular one or more genes because a region need not take into specific account the particular coding segments (exons) of an individual gene.

The term genotyping refers to the determination of the genetic information an individual carries at one or more positions in the genome. For example, genotyping may comprise the determination of which allele or alleles an individual carries for a single SNP or the determination of which allele or alleles an individual carries for a plurality of SNPs.

The term "target sequence", "target nucleic acid" or "target" refers to a nucleic acid of interest. The target sequence may or may not be of biological significance. Typically, though not always, it is the significance of the target sequence which is being studied in a particular experiment. As non-limiting examples, target sequences may include regions of genomic DNA which are believed to contain one or more polymorphic sites, DNA encoding or believed to encode genes or portions of genes of known or unknown function, DNA encoding or believed to encode proteins or portions of proteins

of known or unknown function, DNA encoding or believed to encode regulatory regions such as promoter sequences, splicing signals, polyadenylation signals, etc. The number of sequences to be interrogated can vary, but preferably are from 1, 10, 100, 1000, or 10,000, 100,000 or 1,000,000 target sequences.

The term subset or representative subset refers to a fraction of a genome. The subset may be 0.1, 1, 3, 5, 10, 25, 50 or 75% of the genome. The partitioning of fragments into subsets may be done according to a variety of physical characteristics of individual fragments. For example, fragments may be divided into subsets according to size, according to the particular combination of restriction sites at the ends of the fragment, or based on the presence or absence of one or more particular sequences.

An "array" comprises a support, preferably solid, with nucleic acid probes attached to the support. Preferred arrays typically comprise a plurality of different nucleic acid probes that are coupled to a surface of a substrate in different, known locations. These arrays, also described as "microarrays" or colloquially "chips" have been generally described in the art, for example, U.S. Pat. Nos. 5,143,854, 5,445,934, 5,744,305, 5,677,195, 5,800,992, 6,040,193, 5,424,186 and Fodor et al., *Science*, 251:767-777 (1991). Each of which is incorporated by reference in its entirety for all purposes.

Arrays may generally be produced using a variety of techniques, such as mechanical synthesis methods or light directed synthesis methods that incorporate a combination of photolithographic methods and solid phase synthesis methods. Techniques for the synthesis of these arrays using mechanical synthesis methods are described in, e.g., U.S. Pat. No. 5,384,261, and 6,040,193, which are incorporated herein by reference in their entirety for all purposes. Although a planar array surface is preferred, the array may be fabricated on a surface of virtually any shape or even a multiplicity of surfaces. Arrays may be nucleic acids on beads, gels, polymeric surfaces, fibers such as fiber optics, glass or any other appropriate substrate. (*See* U.S. Patent Nos. 5,770,358, 5,789,162, 5,708,153, 6,040,193 and 5,800,992, which are hereby incorporated by reference in their entirety for all purposes.)

Arrays may be packaged in such a manner as to allow for diagnostic use or can be an all-inclusive device; e.g., U.S. Patent Nos. 5,856,174 and 5,922,591 incorporated in their entirety by reference for all purposes.

Preferred arrays are commercially available from Affymetrix under the brand name GeneChip® and are directed to a variety of purposes, including genotyping and gene expression monitoring for a variety of eukaryotic and prokaryotic species. (*See* Affymetrix Inc., Santa Clara and their website at affymetrix.com.)

Hybridization probes are oligonucleotides capable of binding in a base-specific manner to a complementary strand of nucleic acid. Such probes include peptide nucleic acids, as described in Nielsen et al., *Science* 254, 1497-1500 (1991), and other nucleic acid analogs and nucleic acid mimetics. See US Patent Application No. 08/630,427-filed 4/3/96.

Hybridizations are usually performed under stringent conditions, for example, at a salt concentration of no more than 1 M and a temperature of at least 25°C. For example, conditions of 5X SSPE (750 mM NaCl, 50 mM NaPhosphate, 5 mM EDTA, pH 7.4) and a temperature of 25-30°C are suitable for allele-specific probe hybridizations. For stringent conditions, see, for example, Sambrook, Fritsche and Maniatis. "Molecular Cloning A laboratory Manual" 2nd Ed. Cold Spring Harbor Press (1989) which is hereby incorporated by reference in its entirety for all purposes above.

Polymorphism refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A polymorphic marker or site is the locus at which divergence occurs. Preferred markers have at least two alleles, each occurring at frequency of preferably greater than 1%, and more preferably greater than 10% or 20% of a selected population. A polymorphism may comprise one or more base changes, an insertion, a repeat, or a deletion. A polymorphic locus may be as small as one base pair. Polymorphic markers include restriction fragment length polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements such as Alu. The first identified allelic form is arbitrarily designated as the reference form and other allelic forms are designated as alternative or variant alleles. The allelic form occurring most frequently in a selected population is

sometimes referred to as the wildtype form. Diploid organisms may be homozygous or heterozygous for allelic forms. A diallelic polymorphism has two forms. A triallelic polymorphism has three forms. A polymorphism between two nucleic acids can occur naturally, or be caused by exposure to or contact with chemicals, enzymes, or other agents, or exposure to agents that cause damage to nucleic acids, for example, ultraviolet radiation, mutagens or carcinogens.

Single nucleotide polymorphisms (SNPs) are positions at which two alternative bases occur at appreciable frequency (>1%) in the human population, and are the most common type of human genetic variation. The site is usually preceded by and followed by highly conserved sequences of the allele (e.g., sequences that vary in less than 1/100 or 1/1000 members of the populations).

A single nucleotide polymorphism usually arises due to substitution of one nucleotide for another at the polymorphic site. A transition is the replacement of one purine by another purine or one pyrimidine by another pyrimidine. A transversion is the replacement of a purine by a pyrimidine or vice versa. Single nucleotide polymorphisms can also arise from a deletion of a nucleotide or an insertion of a nucleotide relative to a reference allele.

An individual is not limited to a human being, but may also include other organisms including but not limited to mammals, plants, bacteria or cells derived from any of the above.

In silico digestion is a computer aided simulation of enzymatic digests accomplished by searching a sequence for restriction sites. *In silico* digestion provides for the use of a computer system to model enzymatic reactions in order to determine experimental conditions before conducting any actual experiments. An example of an experiment would be to model digestion of the human genome with specific restriction enzymes to predict the sizes of the resulting restriction fragments.

The Noise Likelihood Factor (NLF) is a measure of the likelihood that a SNP is "present" in the subset of genome that was hybridized to the array. A decreasing NLF score is an indication of increased confidence that the SNP is present in the sample. A score of –6 was used as the cut off for determining significance. The NLF cutoff can be set at a more stringent level of –10 or –15.

(C.) Complexity Management

The present invention provides for novel methods of sample preparation and analysis involving managing or reducing the complexity of a nucleic acid sample, such as genomic DNA, by amplifying a representative subset of the sample. The invention further provides for analysis of the above subset by hybridization to an array which may be specifically designed to interrogate the desired fragments for particular characteristics, such as, for example, the presence or absence of a polymorphism. The invention is particularly useful when combined with other methods of genome analysis. As an example, the present techniques are useful to genotype individuals after polymorphisms have been identified.

One method that has been used to isolate a subset of a genome is to separate fragments according to size by electrophoresis in a gel matrix. The region of the gel containing fragments in the desired size range is then excised and the fragments are purified away from the gel matrix. The SNP consortium (TSC) adopted this approach in their efforts to discover single nucleotide polymorphisms (SNPs) in the human genome. *See*, Altshuler *et al.*, *Science* 407: 513-516 (2000) and The International SNP Map Working Group, *Nature* 409: 928-933 (2001) both of which are herein incorporated by reference in their entirety for all purposes.

The present invention provides methods of complexity management of nucleic acid samples, such as genomic DNA, that can be used as an alternative to separation of fragments by gel electrophoresis and purification of fragments from a gel matrix. Generally, the embodiments include the steps of: fragmenting the nucleic acid by digestion with one or more restriction enzymes or through alternative methods of fragmentation; ligating adaptors to the ends of the fragments; and amplifying a subset of the fragments using amplification conditions that selectively amplify fragments of a desired size range. In a preferred embodiment the amplified sequences are then exposed to an array which may or may not have been specifically designed and manufactured to interrogate the isolated sequences. Design of both the complexity management steps and the arrays may be aided by computer modeling techniques. Generally, the steps of the present invention involve reducing the complexity of a nucleic acid sample using the

disclosed techniques alone or in combination. None of these techniques requires purification of the fragments from a gel matrix.

When interrogating genomes it is often useful to first reduce the complexity of the sample and analyze one or more subsets of the genome. Subsets can be defined by many characteristics of the fragments. In a preferred embodiment of the current invention, the subsets are defined by the size of the fragments. Useful size ranges may be from 100, 200, 400, 700 or 1000 to 500, 800, 1500, 2000, 4000 or 10,000. However, larger size ranges such as 4000, 10,000 or 20,000 to 10,000, 20,000 or 500,000 base pairs may also be useful.

It will be understood by those of skill in the art that a subset will be composed primarily of fragments from the selected size range, but some fragments that are longer or shorter than the selected size range may be present in the amplification product.

The genomic DNA sample of the current invention may be isolated according to methods known in the art, such as PCR, reverse transcription, and the like. It may be obtained from any biological or environmental source, including plant, animal (including human), bacteria, fungi or algae. Any suitable biological sample can be used for assay of genomic DNA. Convenient suitable samples include whole blood, tissue, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair.

In a preferred embodiment of the invention, adaptors are ligated to the ends of the fragments and the fragments are amplified by PCR using one or more primers that are designed to hybridize to sequences in the adaptors. In a particularly preferred embodiment, a single primer or primer pair can be used for amplification.

There are many known methods of amplifying nucleic acid sequences including e.g., PCR. *See,* e.g., *PCR Technology: Principles and Applications for DNA Amplification* (ed. H.A. Erlich, Freeman Press, NY, NY, 1992); *PCR Protocols: A Guide to Methods and Applications* (eds. Innis, et al., Academic Press, San Diego, CA, 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications* 1, 17 (1991); *PCR* (eds. McPherson et al., IRL Press, Oxford); and U.S. Patent 4,683,202, 4,683,195, 4,800,159 4,965,188 and 5,333,675 each of which is incorporated herein by reference in their entireties for all purposes.

PCR is an extremely powerful technique for amplifying specific polynucleotide sequences, including genomic DNA, single-stranded cDNA, and mRNA among others. Various methods of conducting PCR amplification and primer design and construction for PCR amplification will be known to those of skill in the art. Generally, in PCR a double stranded DNA to be amplified is denatured by heating the sample. New DNA synthesis is then primed by hybridizing primers to the target sequence in the presence of DNA polymerase and excess dNTPs. In subsequent cycles, the primers hybridize to the newly synthesized DNA to produce discreet products with the primer sequences at either end. The products accumulate exponentially with each successive round of amplification.

The DNA polymerase used in PCR is often a thermostable polymerase. This allows the enzyme to continue functioning after repeated cycles of heating necessary to denature the double stranded DNA. Polymerases that are useful for PCR include, for example, *Taq* DNA polymerase, *Tth* DNA polymerase, *Tfl* DNA polymerase, *Tma* DNA polymerase, *Tli* DNA polymerase, and *Pfu* DNA polymerase. There are many commercially available modified forms of these enzymes including: AmpliTaq® and AmpliTaq Gold® both available from Applied Biosystems. Many are available with or without a 3- to 5' proofreading exonuclease activity. See, for example, Vent® and Vent® (exo-) available from New England Biolabs.

Other suitable amplification methods include the ligase chain reaction (LCR) (*e.g.,* Wu and Wallace, *Genomics* 4, 560 (1989) and Landegren et al., *Science* 241, 1077 (1988)), transcription amplification (Kwoh et al., *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989)), and self-sustained sequence replication (Guatelli et al., *Proc. Nat. Acad. Sci. USA,* 87, 1874 (1990)) and nucleic acid based sequence amplification (NABSA). (*See,* US patents nos. 5,409,818, 5,554517, and 6,063,603). The latter two amplification methods include isothermal reactions based on isothermal transcription, which produce both single-stranded RNA (ssRNA) and double-stranded DNA (dsDNA) as the amplification products in a ratio of about 30 or 100 to 1, respectively.

When genomic DNA is digested with one or more restriction enzymes the sizes of the fragments are randomly distributed over a broad range. Following adaptor ligation, all of the fragments that have adaptors ligated to both ends will compete equally for

primer binding and amplification in many methods of amplification, regardless of size. The present invention describes methods for biasing amplification toward fragments of a selected size range. Some of the conditions that can be changed to control the fragment size of amplification products, in for example a PCR, are shown in Figure 1. For example short extension times favor amplification of smaller fragments because length of extension is limited by extension time. Longer fragments require longer extension times because shorter extension times can result in termination of the extension product prior to completion. These prematurely terminated products will not serve as templates for subsequent rounds of amplification because they will be lacking the required primer binding site at their 3' end. The average size of the amplification product can also be controlled by varying the length of the adaptor, the sequence of the adaptor and the concentration of primer in the amplification reaction. The size of the amplification product can also be controlled by the addition of varying concentrations of chain terminating nucleotide analogs or an enzyme activity, such as a 3' to 5' exonuclease activity. Because of the geometric nature of PCR amplification, subtle differences in yields that occur in the initial cycles, will result in significant differences in yields in later cycles. (*See*, PCR Primer: A Laboratory Manual, CSHL Press, Eds. Carl Dieffenbach and Gabriela Dveskler, (1995), (Dieffenbach et al.) which is herein incorporated by reference in its entirety for all purposes.)

In a preferred embodiment of the current invention, schematically illustrated in Figure 2, the average length of the amplification product is controlled by ligating adaptors to the fragments that introduce a region of complementarity between the 5' and 3' ends of the fragment. During the primer annealing phase of amplification, self-annealing of the ends of the template will compete with binding of the primer to the template. The probability of self-annealing is proportional to the length of the fragment so the probability that the ends of a shorter fragment will self-anneal is higher than the probability that the ends of a longer fragment will self-anneal. (*See*, Brownie et al, *Nucleic Acids Research*, 25:3235-3241, (1997), which is herein incorporated by reference in its entirety for all purposes.)

In this embodiment, the length of the amplification product is also dependent on the concentration of added primer. Primer binding is competing with self-annealing so

the probability that a primer will bind depends on the concentration of the primer. If, for example, the desired fragment size is short, higher primer concentration should be used. In general, higher primer concentrations favor primer binding and amplification of shorter fragments while decreased primer concentration favors amplification of longer fragments. Preferably the range of primer concentration is from 0.1, 0.3, or 0.5 μM, to 0.5, 1, 2 or 10 μM. Figure 3 shows the effect of increasing primer concentration.

The concentration of salt in the reaction, for example, $MgCl_2$ may also be manipulated to favor amplification of a selected size of fragments. It may be necessary to titrate salt concentration for optimization. The denaturation temperature can also be varied to favor amplification of selected fragments. (*See*, Current Protocols in Molecular Biology, eds. Ausubel et al. (2000), which is herein incorporated by reference in its entirety for all purposes.) For example, denaturation temperatures under 94°C select against amplification of fragments that are GC rich.

Competition between self-annealing and primer annealing can also be regulated by varying the length of the complementarity between the ends of the target sequence in relation to the length of complementarity between the target and the primer. Longer regions of complementarity, for example 30, 40 or 50 to 50, 80 or 100, base pairs between the ends of the fragment favor self-annealing and increased average length of the amplification product. Shorter regions of complementarity, for example 1, 5, 10 or 20 to 5, 10, or 25 base pairs of complementarity favors shorter amplification products. The complementary regions may be at the very ends of the fragments but may also be within 50 or 100 bases of the ends. Increased complementarity between the primer and the adaptor favors primer binding.

Inclusion of chain terminating nucleotides or nucleotide analogs can also be used to regulate the average length of an amplification product. Addition of a chain terminator such as ddATP, ddCTP, ddGTP, ddUTP and ddTTP results in the termination of extension whenever one of these is incorporated into an extending strand. The concentration of a chain terminator relative to its corresponding dNTP determines the frequency of termination. The result is that longer fragments are more likely to be prematurely terminated than shorter fragments. (*See*, Current Protocols in Molecular Biology, eds. Ausubel et al. (2000), which is herein incorporated by reference in its

entirety for all purposes.) Useful concentrations of a chain terminator are, for example, dNTP:ddNTP equals 100:1 or 1000:1. The ratio of dNTP:ddNTP can be varied depending on the desired average length of the products and the relative binding affinities of the enzyme for the dNTP and ddNTP.

Inclusion of a 3' to 5' exonuclease activity favors amplification of long DNA by removing nucleotide misincorporations and preventing premature termination of strand synthesis. Conversely, absence of 3' to 5' exonuclease favors amplification of smaller fragments because of misincorporation of nucleotides leading to premature termination of strand synthesis. Many of the thermophilic DNA polymerases available for PCR are commercially available with or without 3' to 5' exonuclease activity, for example Vent and Vent (exo-) both available from New England Biolabs. Other components of the reaction can also be manipulated to effect the size of the amplification products by modifying polymerase fidelity. (*See, for example*, PCR Strategies, eds. Innis et al, Academic Press (1995), (Innis et al.),which is herein incorporated by reference for all purposes).

In one embodiment, shorter amplification products are selected by decreasing the extension and annealing times of the amplification reaction to, for example, from 2, 5, 10 or 20 to 5, 10 or 30 seconds, resulting in preferential amplification of shorter restriction fragments, because the likelihood of completing primer extension on longer fragments is less than on shorter fragments.

Another embodiment further comprises the step of diluting the product of the first round of size selective amplification and subjecting it to a second round of size selective amplification. The second round of amplification further enriches for smaller fragments.

In one embodiment, the restriction digest is further fractionated prior to PCR amplification by applying the sample to a gel exclusion column. For example, to exclude the shortest fragments from the amplification the restriction digest can be passed over a column that selectively retains smaller fragments, for example fragments under 400 base pairs. The larger fragments, over 400 base pairs, can be recovered in the void volume. The fragments in the void volume would then have adaptors ligated to them followed by PCR amplification. Because the shortest fragments in the PCR would be approximately

400 base pairs, the resulting PCR products should be in a size range starting at 400 base pairs.

As those of skill in the art will appreciate, after amplification, the resulting sequences may be further analyzed using any known method including sequencing, HPLC, hybridization analysis, cloning, labeling, etc.

The materials for use in the present invention are ideally suited for the preparation of a kit suitable for obtaining a subset of a genome. Such a kit may comprise various reagents utilized in the methods, preferably in concentrated form. The reagents of this kit may comprise, but are not limited to, buffer, appropriate nucleotide triphosphates, appropriate dideoxynucleotide triphosphates, reverse transcriptases, nucleases, restriction enzymes, adaptors, ligases, DNA polymerases, primers and instructions for the use of the kit.

(D.) Designing an Array to Interrogate Size Selected SNPs

In a particularly preferred embodiment the current invention is combined with *in silico* digestion techniques to predict the SNPs that will be present when a genome is digested with a particular enzyme or enzymes and fragments of a particular size are amplified. In Figure 4, a computer is first used to locate a SNP from the public database provided by The SNP Consortium (TSC), (available at http://snp.cshl.org/ last visited 7/25/2001) in the public database of the sequence of the human genome, available in GenBank (See, www.ncbi.nlm.nih.gov, last visited 7/25/2001). The computer is then used to predict the, for example, *Bgl*II restriction sites upstream and downstream of a given SNP. For example, in Figure 4 TSC SNP ID 10034 has a *Bgl*II site at upstream position 49050 and a downstream *Bgl*II site at position 52100. Given these restriction sites the computer would further predict that SNP 10034 would be contained on a 3050 base pair fragment when genomic DNA was digested with *Bgl*II.

The SNPs and corresponding fragment sizes could be further separated by computer into subsets according to fragment size. In this way a computer could be used to identify all of the SNPs that are predicted to be found on fragments that are between 400 and 800 base pairs in length when a sample DNA is digested with a given one or more enzymes. The information can also be used to design arrays to interrogate those

SNPs predicted to be present in a particular size fraction resulting from a particular digestion method.

In Figure 5, *in silico* digestion was used to predict restriction fragment lengths for the more than 800,000 SNPs in the TSC database and to identify those SNP containing fragments between 400 and 800 base pairs. For example, when human genomic DNA is digested with *Eco*RI, 32,908 SNPs from the TSC database are predicted to be found on fragments between 400 and 800 base pairs. More than 120,000 of the TSC SNPs are found on fragments between 400 and 800 base pairs when genomic DNA is digested with *Eco*RI, *Xba*I, *Pst*I and *Bgl*II.

Figure 6 shows a schematic of how the current invention further provides methods to combine *in silico* prediction of the size of SNP containing fragments with methods of size selection by PCR to design genotyping assays and arrays for genotyping. In the figure the selected size range is 400 to 800 base pairs but other size ranges could also be used, for example, 100, 200, 500, 700, 1,500, 5,000 or 10,000 to 500, 700, 1,000, 2,000, 3,000, 10,000 or 20,000 base pairs may also be useful size ranges.

As shown in Figure 6, in this embodiment of the current invention an array is designed to interrogate the SNPs that are predicted to be found in a size fraction resulting from digestion of the first nucleic acid sample with one or more particular restriction enzymes. For example, a computer may be used to search the sequence of a genome to identify all recognition sites for the restriction enzyme, *Eco*RI. The computer can then be used to predict the size of all restriction fragments resulting from an *Eco*RI digestion and to identify those fragments that contain a known or suspected SNP. The computer may then be used to identify the group of SNPs that are predicted to be found on fragments of, for example, 400-800 base pairs, when genomic DNA is digested with *Eco*RI. An array may then be designed to interrogate that subset of SNPs that are found on *Eco*RI fragments of 400-800 base pairs.

The design of the array may be further refined by adding additional information about each SNP. For example, subsequently obtained empirical data about a particular SNP may indicate that fewer probes are necessary to determine the presence of a given allele. SNPs that prove to be of particular biological importance may be added and SNPs that are subsequently shown to be of little or no biological importance can be removed.

Arrays will preferably be designed to interrogate 100, 500, 1000, 5000, 10,000, 50,000 or 100,000 different SNPs. For example, an array may be designed to recognize a group of SNPs predicted to be present on 400-800 base pair *EcoRI* fragments, a collection of SNPs predicted to be present on 400-800 base pair *Bgl*II fragments, a collection of SNPs predicted to be present on 400-800 base pair *XbaI* fragments, and a collection of SNPs predicted to be present on 400-800 base pair *Hind*III fragments. One or more PCR products, that differ in the restriction enzyme used for fragmentation, could be pooled prior to hybridization to increase the complexity of the sample.

In one embodiment of the invention a single size selected amplification product is suitable for hybridization to many different arrays. For example, a single method of fragmentation and amplification that is suitable for hybridization to an array designed to interrogate SNPs contained on 400-800 base pair *Eco*RI would also be suitable for hybridization to an array designed to interrogate SNPs contained on 400-800 base pair *Bam*HI fragments. This would introduce consistency and reproducibility to sample preparation methods.

The methods of the current invention can also be adapted to take advantage of a bias in the publicly available SNP database. Many of these SNPs were identified from genomic libraries constructed by digesting genomic DNA with specific enzymes, for example one group identified SNPs on *Bgl*II, *Hind*III, *Xba*I and *Eco*RI fragments. *See*, Altshuler *et al.*, *Science* 407: 513-516 (2000) and The International SNP Map Working Group, *Nature* 409: 928-933 (2001). The resulting fragments were separated by size using agarose gel electrophoresis. A slice of the gel corresponding to a size range of 500-600bp was excised from the gel and the fragments purified from the slice were cloned and sequenced. As a result many of the SNPs identified by the TSC are located on fragments that are smaller than 1kb when genomic DNA is digested with *Bgl*II, *Hind*III, *Xba*I or *Eco*RI. In one embodiment, the present invention provides methods that take advantage of this bias in the publicly available SNPs by digesting genomic DNA with one or more of the following enzymes: *Bgl*II, *Hind*III, *Xba*I and *Eco*RI prior to size selective PCR.

METHODS OF USE

The methods of the presently claimed invention can be used for a wide variety of applications. Any analysis of genomic DNA may be benefited by a reproducible method of complexity management. Furthermore, the methods and enriched fragments of the presently claimed invention are particularly well suited for study and characterization of extremely large regions of genomic DNA.

In a preferred embodiment, the methods of the presently claimed invention are used for SNP discovery and to genotype individuals. For example, any of the procedures described above, alone or in combination, could be used to isolate the SNPs present in one or more specific regions of genomic DNA. Selection probes could be designed and manufactured to be used in combination with the methods of the invention to amplify only those fragments containing regions of interest, for example a region known to contain a SNP. Arrays could be designed and manufactured on a large scale basis to interrogate only those fragments containing the regions of interest. Thereafter, a sample from one or more individuals would be obtained and prepared using the same techniques which were used to prepare the selection probes or to design the array. Each sample can then be hybridized to an array and the hybridization pattern can be analyzed to determine the genotype of each individual or a population of individuals. Methods of use for polymorphisms and SNP discovery can be found in, for example, co-pending US application Nos. 08/813,159 and 09/428,350 which are herein incorporated by reference in their entirety for all purposes).

Correlation of Polymorphisms with Phenotypic Traits

Most human sequence variation is attributable to or correlated with SNPs, with the rest attributable to insertions or deletions of one or more bases, repeat length polymorphisms and rearrangements. On average, SNPs occur every 1,000-2,000 bases when two human chromosomes are compared. (*See*, The International SNP Map Working Group, *Science* 409: 928-933 (2001) incorporated herein by reference in its entirety for all purposes.) Human diversity is limited not only by the number of SNPs occurring in the genome but further by the observation that specific combinations of alleles are found at closely linked sites.

Correlation of individual polymorphisms or groups of polymorphisms with phenotypic characteristics is a valuable tool in the effort to identify DNA variation that contributes to population variation in phenotypic traits. Phenotypic traits include physical characteristics, risk for disease, and response to the environment. Polymorphisms that correlate with disease are particularly interesting because they represent mechanisms to accurately diagnose disease and targets for drug treatment. Hundreds of human diseases have already been correlated with individual polymorphisms but there are many diseases that are known to have an, as yet unidentified, genetic component and many diseases for which a component is or may be genetic.

Many diseases may correlate with multiple genetic changes making identification of the polymorphisms associated with a given disease more difficult. One approach to overcome this difficulty is to systematically explore the limited set of common gene variants for association with disease.

To identify correlation between one or more alleles and one or more phenotypic traits, individuals are tested for the presence or absence of polymorphic markers or marker sets and for the phenotypic trait or traits of interest. The presence or absence of a set of polymorphisms is compared for individuals who exhibit a particular trait and individuals who exhibit lack of the particular trait to determine if the presence or absence of a particular allele is associated with the trait of interest. For example, it might be found that the presence of allele A1 at polymorphism A correlates with heart disease. As an example of a correlation between a phenotypic trait and more than one polymorphism, it might be found that allele A1 at polymorphism A and allele B1 at polymorphism B correlate with a phenotypic trait of interest.

Diagnosis of Disease and Predisposition to Disease

Markers or groups of markers that correlate with the symptoms or occurrence of disease can be used to diagnose disease or predisposition to disease without regard to phenotypic manifestation. To diagnose disease or predisposition to disease, individuals are tested for the presence or absence of polymorphic markers or marker sets that correlate with one or more diseases. If, for example, the presence of allele A1 at

polymorphism A correlates with coronary artery disease then individuals with allele A1 at polymorphism A may be at an increased risk for the condition.

Individuals can be tested before symptoms of the disease develop. Infants, for example, can be tested for genetic diseases such as phenylketonuria at birth. Individuals of any age could be tested to determine risk profiles for the occurrence of future disease. Often early diagnosis can lead to more effective treatment and prevention of disease through dietary, behavior or pharmaceutical interventions. Individuals can also be tested to determine carrier status for genetic disorders. Potential parents can use this information to make family planning decisions.

Individuals who develop symptoms of disease that are consistent with more than one diagnosis can be tested to make a more accurate diagnosis. If, for example, symptom S is consistent with diseases X, Y or Z but allele A1 at polymorphism A correlates with disease X but not with diseases Y or Z an individual with symptom S is tested for the presence or absence of allele A1 at polymorphism A. Presence of allele A1 at polymorphism A is consistent with a diagnosis of disease X. Genetic expression information discovered through the use of arrays has been used to determine the specific type of cancer a particular patient has. (*See*, Golub et al. Science 286: 531-537 (2001) hereby incorporated by reference in its entirety for all purposes.)

Pharmacogenomics

Pharmacogenomics refers to the study of how genes affect response to drugs. There is great heterogeneity in the way individuals respond to medications, in terms of both host toxicity and treatment efficacy. There are many causes of this variability, including: severity of the disease being treated; drug interactions; and the individuals age and nutritional status. Despite the importance of these clinical variables, inherited differences in the form of genetic polymorphisms can have an even greater influence on the efficacy and toxicity of medications. Genetic polymorphisms in drug-metabolizing enzymes, transporters, receptors, and other drug targets have been linked to interindividual differences in the efficacy and toxicity of many medications. (*See*, Evans and Relling, *Science* 286: 487-491 (2001) which is herein incorporated by reference for all purposes).

An individual patient has an inherited ability to metabolize, eliminate and respond to specific drugs. Correlation of polymorphisms with pharmacogenomic traits identifies those polymorphisms that impact drug toxicity and treatment efficacy. This information can be used by doctors to determine what course of medicine is best for a particular patient and by pharmaceutical companies to develop new drugs that target a particular disease or particular individuals within the population, while decreasing the likelihood of adverse affects. Drugs can be targeted to groups of individuals who carry a specific allele or group of alleles. For example, individuals who carry allele A1 at polymorphism A may respond best to medication X while individuals who carry allele A2 respond best to medication Y. A trait may be the result of a single polymorphism but will often be determined by the interplay of several genes.

In addition some drugs that are highly effective for a large percentage of the population, prove dangerous or even lethal for a very small percentage of the population. These drugs typically are not available to anyone. Pharmacogenomics can be used to correlate a specific genotype with an adverse drug response. If pharmaceutical companies and physicians can accurately identify those patients who would suffer adverse responses to a particular drug, the drug can be made available on a limited basis to those who would benefit from the drug.

Similarly, some medications may be highly effective for only a very small percentage of the population while proving only slightly effective or even ineffective to a large percentage of patients. Pharmacogenomics allows pharamaceutical companies to predict which patients would be the ideal candidate for a particular drug, thereby dramatically reducing failure rates and providing greater incentive to companies to continue to conduct research into those drugs.

Determination of Relatedness

There are many circumstances where relatedness between individuals is the subject of genotype analysis and the present invention can be applied to these procedures. Paternity testing is commonly used to establish a biological relationship between a child and the putative father of that child. Genetic material from the child can be analyzed for occurrence of polymorphisms and compared to a similar analysis of the putative father's

genetic material. Determination of relatedness is not limited to the relationship between father and child but can also be done to determine the relatedness between mother and child, (see e.g. Staub et al.,U.S. Pat. No.6,187,540) or more broadly, to determine how related one individual is to another, for example, between races or species or between individuals from geographically separated populations, (see for example H. Kaessmann, et al. *Nature Genet.* 22, 78 (1999)).

Forensics

The capacity to identify a distinguishing or unique set of forensic markers in an individual is useful for forensic analysis. For example, one can determine whether a blood sample from a suspect matches a blood or other tissue sample from a crime scene by determining whether the set of polymorphic forms occupying selected polymorphic sites is the same in the suspect and the sample. If the set of polymorphic markers does not match between a suspect and a sample, it can be concluded (barring experimental error) that the suspect was not the source of the sample. If the set of markers does match, one can conclude that the DNA from the suspect is consistent with that found at the crime scene. If frequencies of the polymorphic forms at the loci tested have been determined (e.g., by analysis of a suitable population of individuals), one can perform a statistical analysis to determine the probability that a match of suspect and crime scene sample would occur by chance. A similar comparison of markers can be used to identify an individual's remains. For example the U.S. armed forces collect and archive a tissue sample for each service member. If unidentified human remains are suspected to be those of an individual a sample from the remains can be analyzed for markers and compared to the markers present in the tissue sample initially collected from that individual.

Marker Assisted Breeding

Genetic markers can assist breeders in the understanding, selecting and managing of the genetic complexity of animals and plants. Agriculture industry, for example, has a great deal of incentive to try to produce crops with desirable traits (high yield, disease resistance, taste, smell, color, texture, etc.) as consumer demand increases and

expectations change. However, many traits, even when the molecular mechanisms are known, are too difficult or costly to monitor during production. Readily detectable polymorphisms which are in close physical proximity to the desired genes can be used as a proxy to determine whether the desired trait is present or not in a particular organism. This provides for an efficient screening tool which can accelerate the selective breeding process.

EXAMPLES

Example 1: PCR with Single Adaptor Sequence:

Reagent preparation:

For adaptor and primer preparation dry oligonucleotides were dissolved in 50% glycerol, 50% TE for a final concentration of 100μM and stored at -20°C. Adaptors were made by mixing 100 μl of 50% glycerol, 50% TE , 4 μl 5M NaCl, 50 μl of top strand and 50 μl of bottom strand; heating the mixture at 95°C for 10 minutes in a heating block. The mixtures were then allowed to cool slowly to at least 50°C by shutting off the heat block. The adaptors were used at room temperature. Primers for PCR were diluted to 10 μM with TE.

Genomic DNA was thawed and mixed gently by inverting or tipping the tube, without vortexing. The DNA was pipetted slowly to avoid shearing. If more than 5 μl was needed a P-200 tip that had been cut to enlarge the pore was used. The DNA was then diluted to 0.1 μg/ μl with TE and unused sample was stored at 4°C.

Target preparation.

5 μl of 0.1 μg/ μl human genomic DNA was digested in a 20 μl reaction volume with 20 units restriction enzyme, in 1X RE buffer (NEB) and 1μg/ μl BSA for 2 hours at 37°C. After 1 hour of digestion the reaction was mixed and spun. To stop the reaction the enzyme was inactivated by heating the reaction to 70°C for 20 minutes.

The 20 μl restriction digest was then mixed with 0.5 μl of 25 μM adaptor, 2 μl 100 mM DTT, 2.5 μl 10 mM ATP, and 0.25 μl 2000 units/ μl ligase. The reaction was incubated at 16°C for 2 hours. To stop the reaction 75 μl of TE was added and the reaction was incubated at 95°C for 5 minutes or 70°C for 20 minutes.

PCR test and primer titration

For PCR 2 µl of the ligation reaction was amplified in a 100 µl reaction with concentrations of primer varying from 0.4 to 0.8 µM, 250 µM dNTPs, 2 mM MgCl2, and 5 units TaqGold polymerase. The program was 95°C for 1 minute, 20 cycles of 95°C for 20 sec, 58°C for 20 sec, and 72°C for 20 sec, 25 cycles of 95°C for 20 sec, 55°C for 20 sec, and 72°C for 20 sec, with a final incubation at 72°C for 5 minutes.

An aliquot of the PCR product was analyzed on a gel to confirm that the products had an average size of 400-700 base pairs. Distinct banding patterns were observed for the *Eco*RI and the *Xba*I samples, the presence of the patterns may be a rapid method to assay the quality of a PCR product.

Fragmentation and Labeling

PCR reactions were scaled up to 500 µl and run as above. The product was concentrated to about 40 µl using a filter unit or a Qiagen PCR clean up kit. The PCR products were fragmented with DNase as follows: the concentrated PCR product, about 5 ug DNA, was brought to 40 µl with water and mixed with 0.04 units DNase, buffer and BSA in a total volume of 55 µl. The reaction was incubated for 20 minutes at 37°C then 95°C for 10 minutes.

DNA was labeled by mixing 50 µl of the DNase digestion with 10 µM b-ddNTP, in TDT buffer in a total volume of 99 µl. The mixture was denatured at 95°C for 5 min and cooled to 25°C before adding 1 µl of TdT enzyme, 15-20 units/ µl, and incubated at 37°C for 2 hours to overnight, followed by heat inactivation at 95°C for 10 minutes. Alternatively 2 µl of TdT can be used and the incubation shortened to 1-2 hours at 37°C.

Hybridization

Standard procedures were used for hybridization, washing, scanning and data analysis. Figure 7 shows the results of hybridization of PCR products to an array designed to detect the presence or absence of a given SNP containing target in a sample. Approximately 13,000 SNPs can be assayed on the array. In figure 7A the first nucleic acid was digested with *Bgl*II and the desired size range was 400-700 base pairs. The array is expected to interrogate 3971 SNPs from this subset and 3321 or 83.6% of those were detected with a noise likelihood factor (NLF) of less than -6. As a negative control,

the subset of SNPs contained on *Eco*RI fragments of 400 to 700 base pairs was also included and only 26.4% were detected. Figure 7B and C show results for digestion of the first nucleic acid with *Eco*RI (Fig. 7B) or *Xba*I (Fig. 7C).

Example 2: PCR with Short Extension and Annealing Times and Two Rounds of Amplification.

Figure 8 shows a schematic representation of amplification with two rounds of PCR using short annealing and extension times. For example, 250 ng of Human genomic DNA in TE was digested with 20 units *Bgl*II in a 25 µl reaction volume at 37 °C overnight. The digested DNA was purified using a QIAquick Purification kit (Qiagen) according to the manufacturer's instructions and eluted with 33 µl water. An adaptor, top strand: 5'-GATCAGGCGTCTGTCGTGCTCATAA-3' (SEQ ID NO 1) and bottom strand: 5'- ATTATGAGCACGACAGACGCCT-3' (SEQ ID NO 2), was ligated to the DNA using T4 DNA Ligase at 16 °C for 2 hours. The reaction was stopped by heating the sample to 70 °C for 5 min.

The DNA was then amplified with a first round of PCR in PCR Buffer II with 250 nM primer, 5'-TTATGAGCACGACAGACGCCTGATCT-3' (SEQ ID NO 3), 200 µM dNTPs, 2.5 mM MgCl2, with 5 U AmpliTaq Gold polymerase (Perkin Elmer) in a total reaction volume of 60 µl in 0.2 ml thin wall PCR tubes. PCR was done in a PE 2400 starting with 9 min at 95 °C followed by 25 cycles of 30 sec at 94 °C, 2 sec at 60 °C, and 2 sec at 72 °C. The mixture was finally incubated for 5 min at 72 °C.

For the second round of PCR, 10 µl of the first PCR reaction was diluted into a 100 µl reaction. The concentration of reaction components was the same except 10 U of AmpliTaq Gold polymerase was used. The PCR cycles were the same except 40 cycles were used. The PCR products were purified with QIAquick PCR Purification Kit (Qiagen) according to the manufacturer's instructions and eluted with 33 µl TE. Most of the product following the second round of PCR was in the range of 400 to 1000 base pairs.

We then fragmented the DNA by incubating 30 µl of the purified DNA with 1.1U DNase I (Promega) at 37 °C for 15 min in a 45 µl reaction volume also containing 10

mM Tris-Acetate (pH 7.5), 10 mM magnesium acetate, 50 mM potassium acetate and 1 mM DTT. The reaction was stopped by heating the sample to 95 °C for 15 min.

The sample was then labeled by adding 50 U of terminal transferase and 30 μM of biotin-N^6-ddATP (Dupont-NEN) followed by incubation at 37 °C for 60 min, and heat inactivation at 95 °C for 5 min.

We then hybridized the labeled DNA to an array designed to detect the presence or absence of approximately 13,000 selected SNPs. The hybridization mixture containing 3 M tetraethylamonium chloride, 10 mM MES (pH 6.0), 0.01% Tween-20, 100 μg bovine serum albumin, 1.0 ug COT-1 and 200 pM control oligomer at 44°C for 16 hours.

The array was then washed with 6x SSPE, 0.01% Tween-20 at 25°C then 0.6x SSPE at 40°C. We first stained with staining solution (55 mM MES (pH 6.5), 1.85 M NaCl, 10 μg/ml streptavidin R-phycoerythrin, 2 mg/ml acetylated BSA, 0.1% Tween-20) at 40°C for 15 min. Then we washed with 6x SSPE (0.9 M NaCl, 60 mM NaH_2PO_4 (pH 7.4), 6 mM EDTA , 0.005 % Triton-100) on a fluidics station (Affymetrix) 10 times at 22°C. We then conducted Anti-streptavidin antibody staining at 40°C for 30 min with antibody solution (10 mM MES (pH 6.5), 1 M NaCl, 2 mg anti-streptavidin antibody (Vector), 0.5 mg/ml acetylated BSA, 0.01% Triton-100). We then stained again with staining solution for 15 min followed by 6x SSPE washing as in the previous steps. The arrays were scanned with an Agilent chip scanner at 570 nm.

Figure 9 shows the results from the hybridization and analysis of selected groups of SNPs. The analysis is similar to that of figure 7, when the genomic DNA is digested with BglII the array is predicted to interrogate the presence of 3971 SNPs from the 400 to 700 base pair subset and 2788 of those SNPs were detected with an NLF less than –6. Figure 9B and C show similar analysis for genomic DNA digested with *Eco*RI (Fig. 9B) or *Xba*I (Fig. 9C).

CONCLUSION

From the foregoing it can be seen that the present invention provides a flexible and scalable method for analyzing complex samples of DNA, such as genomic DNA. These methods are not limited to any particular type of nucleic acid sample: plant,

bacterial, animal (including human) total genome DNA, RNA, cDNA and the like may be analyzed using some or all of the methods disclosed in this invention. This invention provides a powerful tool for analysis of complex nucleic acid samples. From experiment design to isolation of desired fragments and hybridization to an appropriate array, the above invention provides for fast, efficient and inexpensive methods of complex nucleic acid analysis.

All publications and patent applications cited above are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent application were specifically and individually indicated to be so incorporated by reference. Although the present invention has been described in some detail by way of illustration and example for purposes of clarity and understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims.